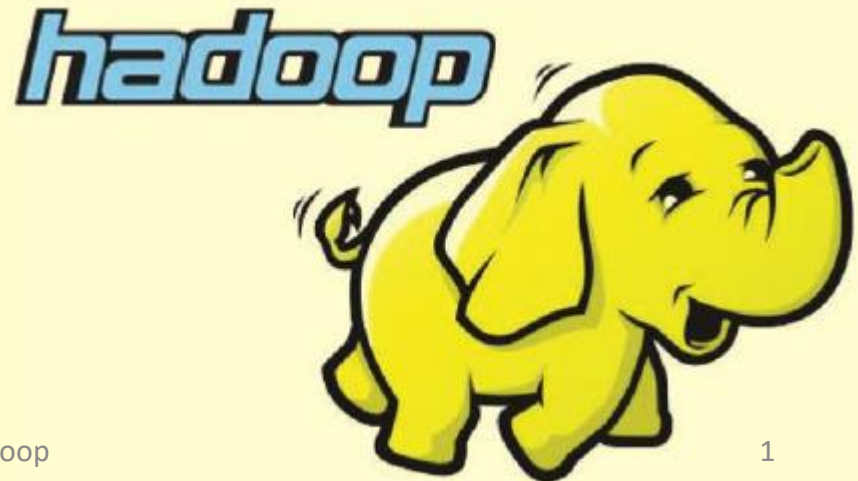
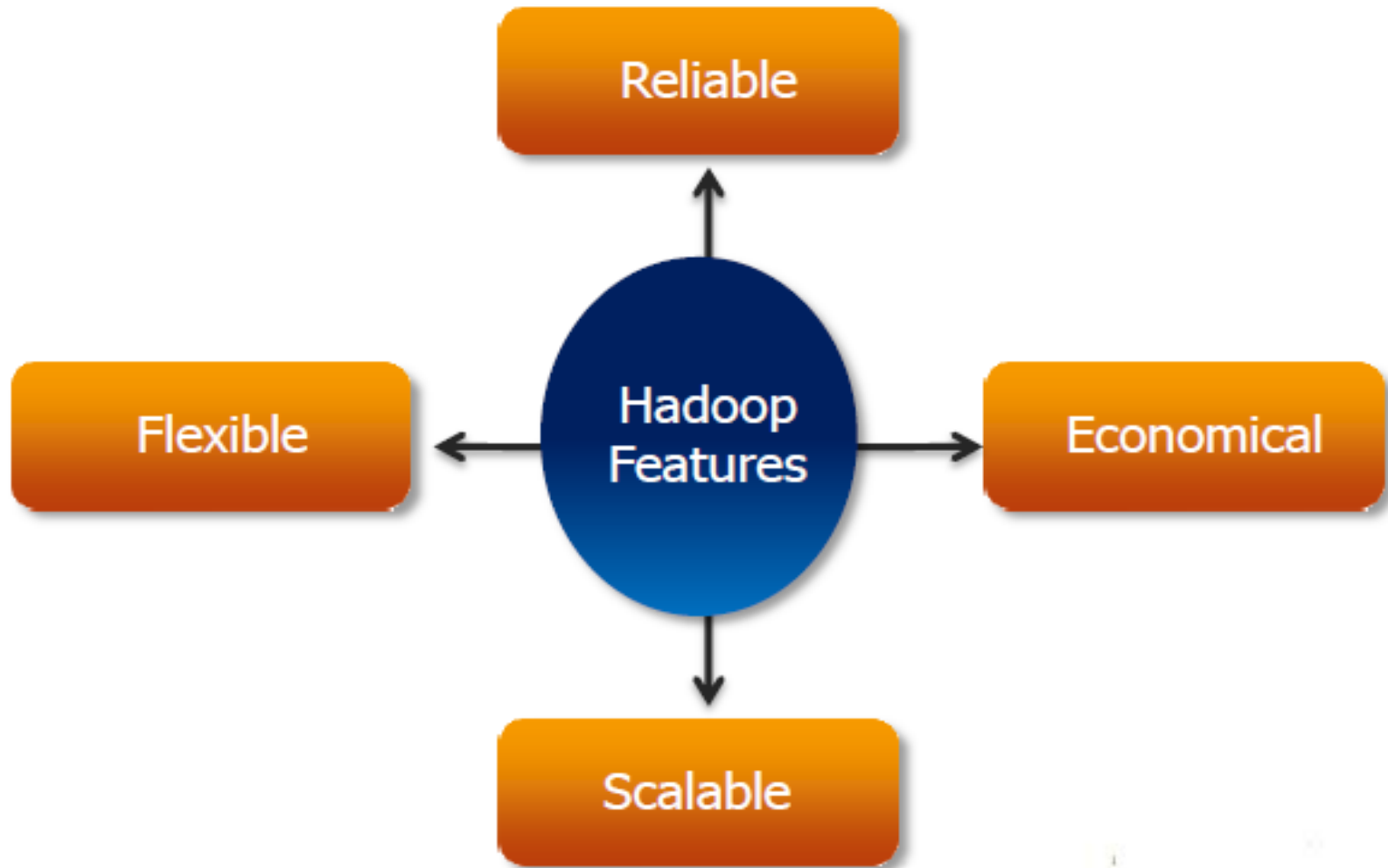


# What is Hadoop ?

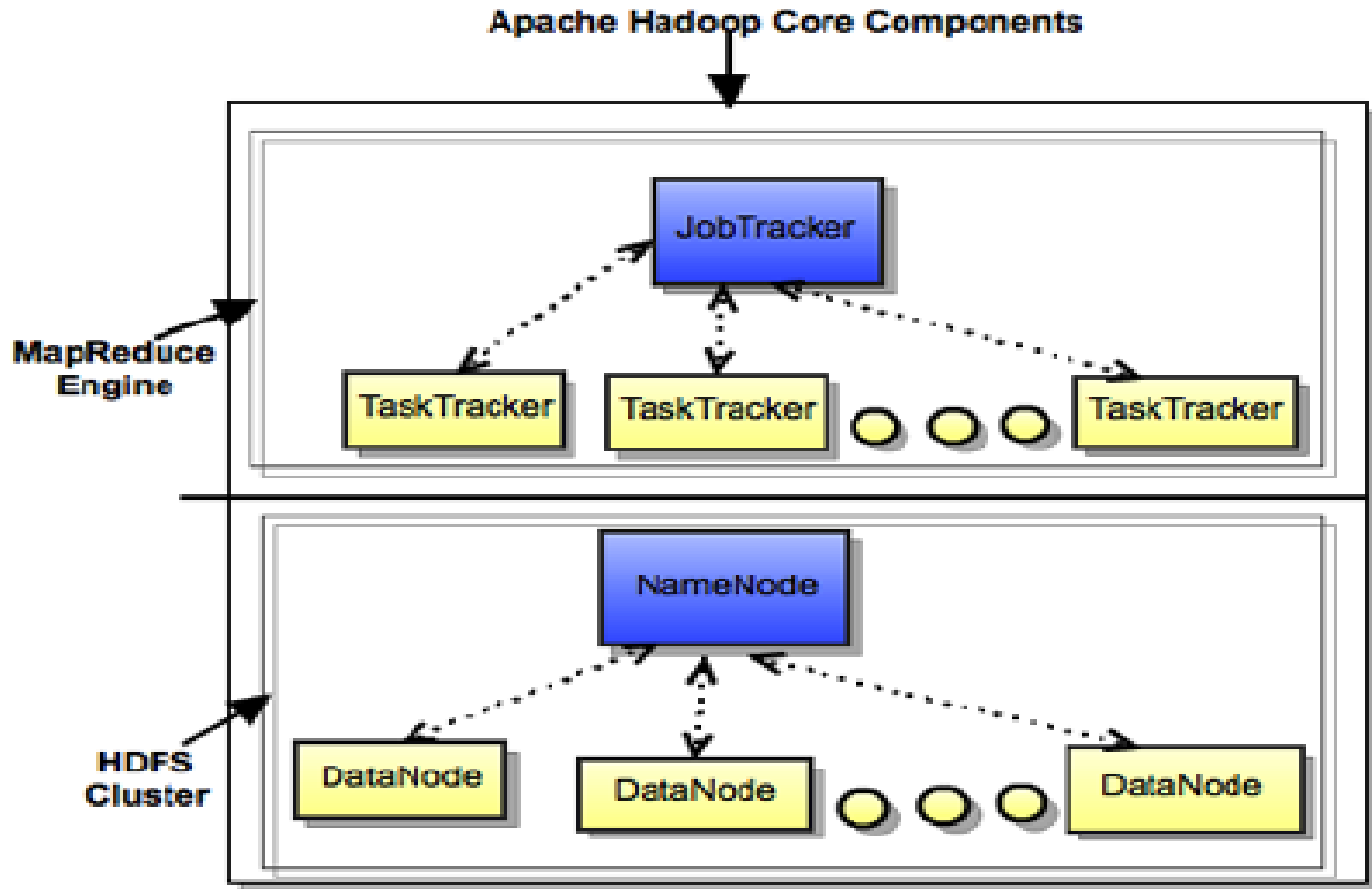
- Apache Hadoop is a **framework** that allows for the **distributed processing of large data sets** across **clusters of commodity computers** using a **simple programming** model.
- **Open-source** Data Management with scale-out storage & distributed pr



# Hadoop – Key Characteristics



# Hadoop Core Components



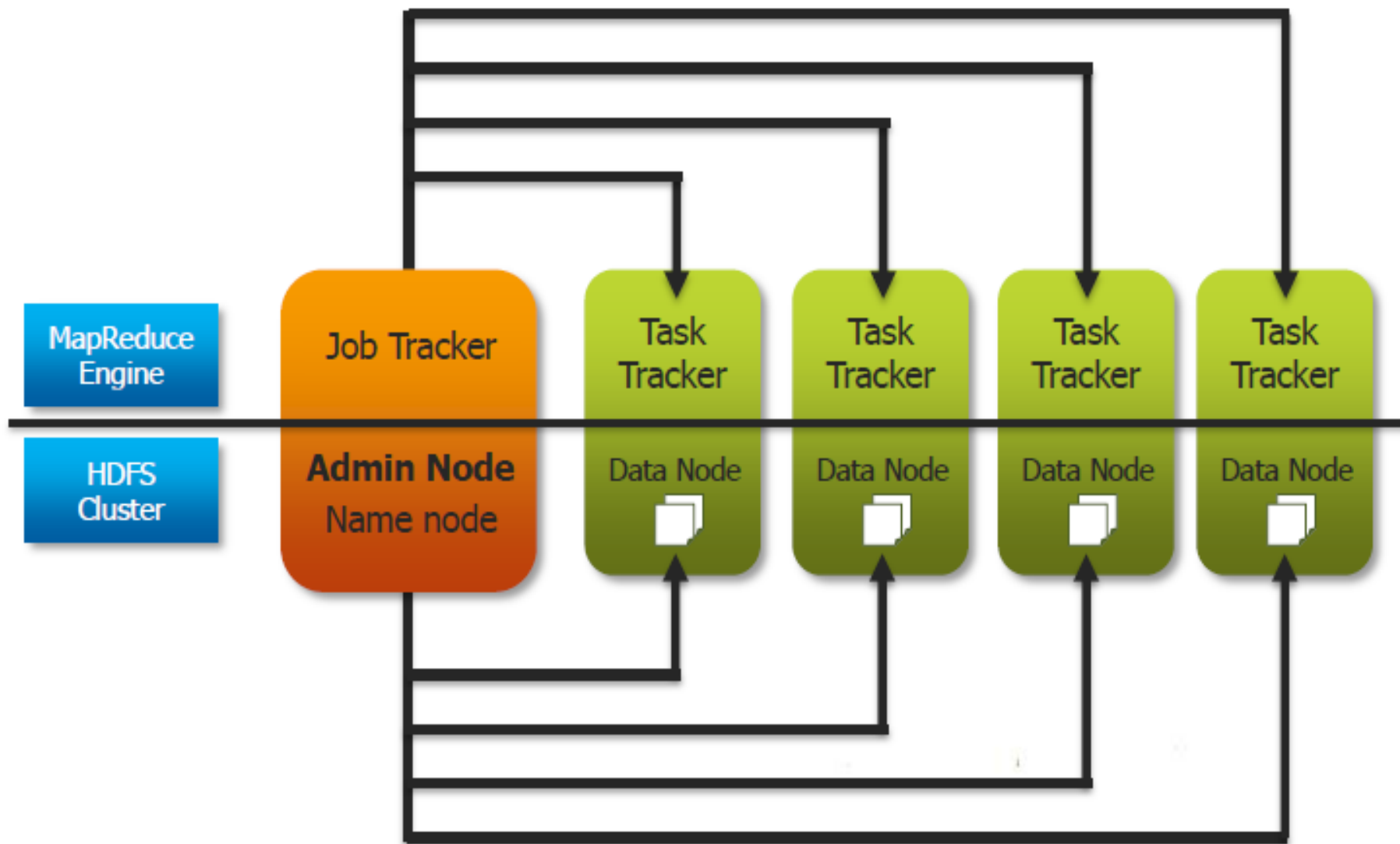
# Hadoop Core Components

**Hadoop is a system for large scale data processing.**

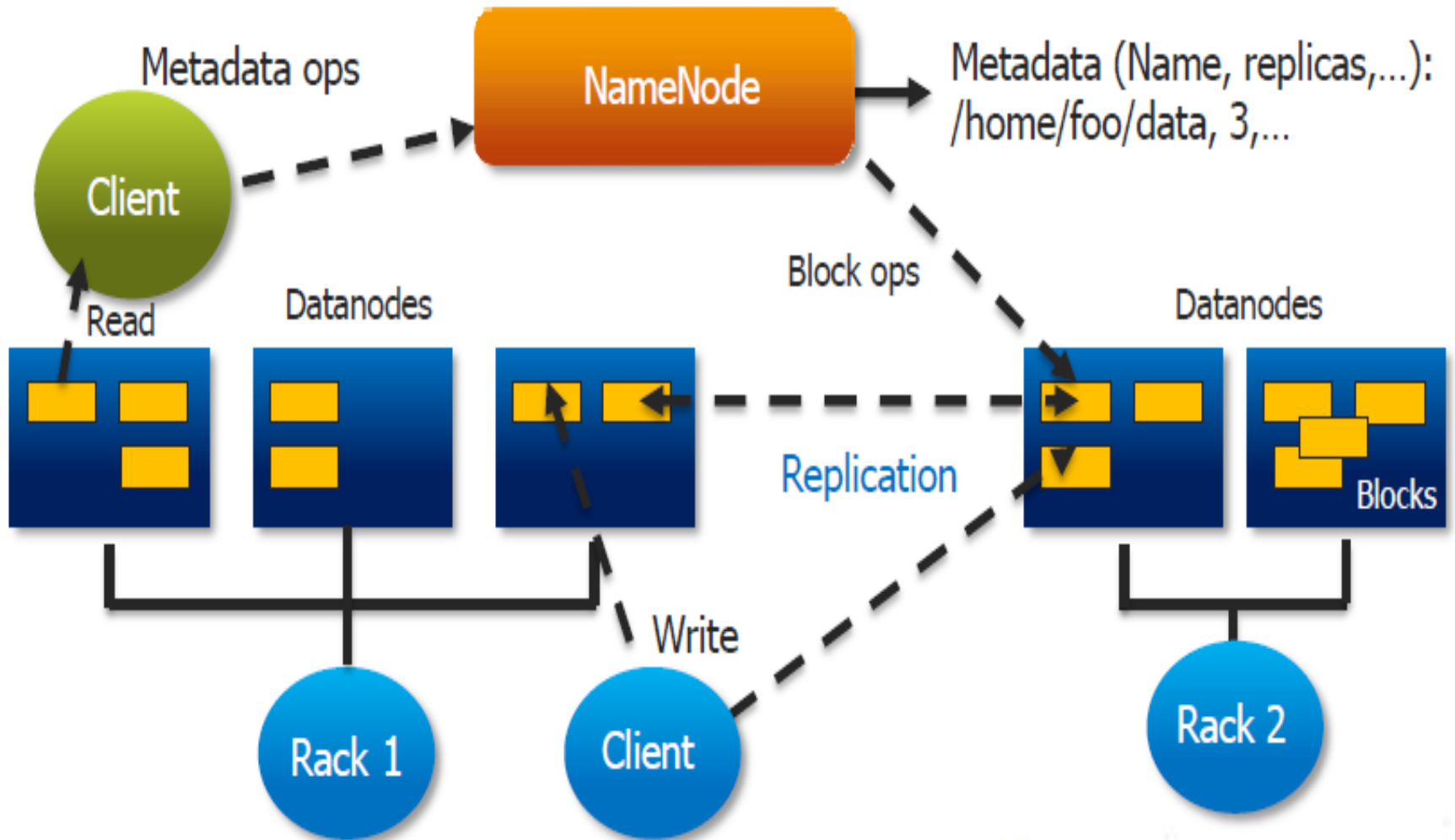
**It has two main components:**

- ✓ **HDFS – Hadoop Distributed File System (Storage)**
  - ✓ Distributed across “nodes”
  - ✓ Natively redundant
  - ✓ NameNode tracks locations.
- ✓ **MapReduce (Processing)**
  - ✓ Splits a task across processors
  - ✓ “near” the data & assembles results
  - ✓ Self-Healing, High Bandwidth
  - ✓ Clustered storage
  - ✓ JobTracker manages the TaskTrackers

# Hadoop Core Components (cont..)



# HDFS Architecture



# Main Components of HDFS

## ✓ NameNode:

- ✓ master of the system
- ✓ maintains and manages the blocks which are present on the DataNodes

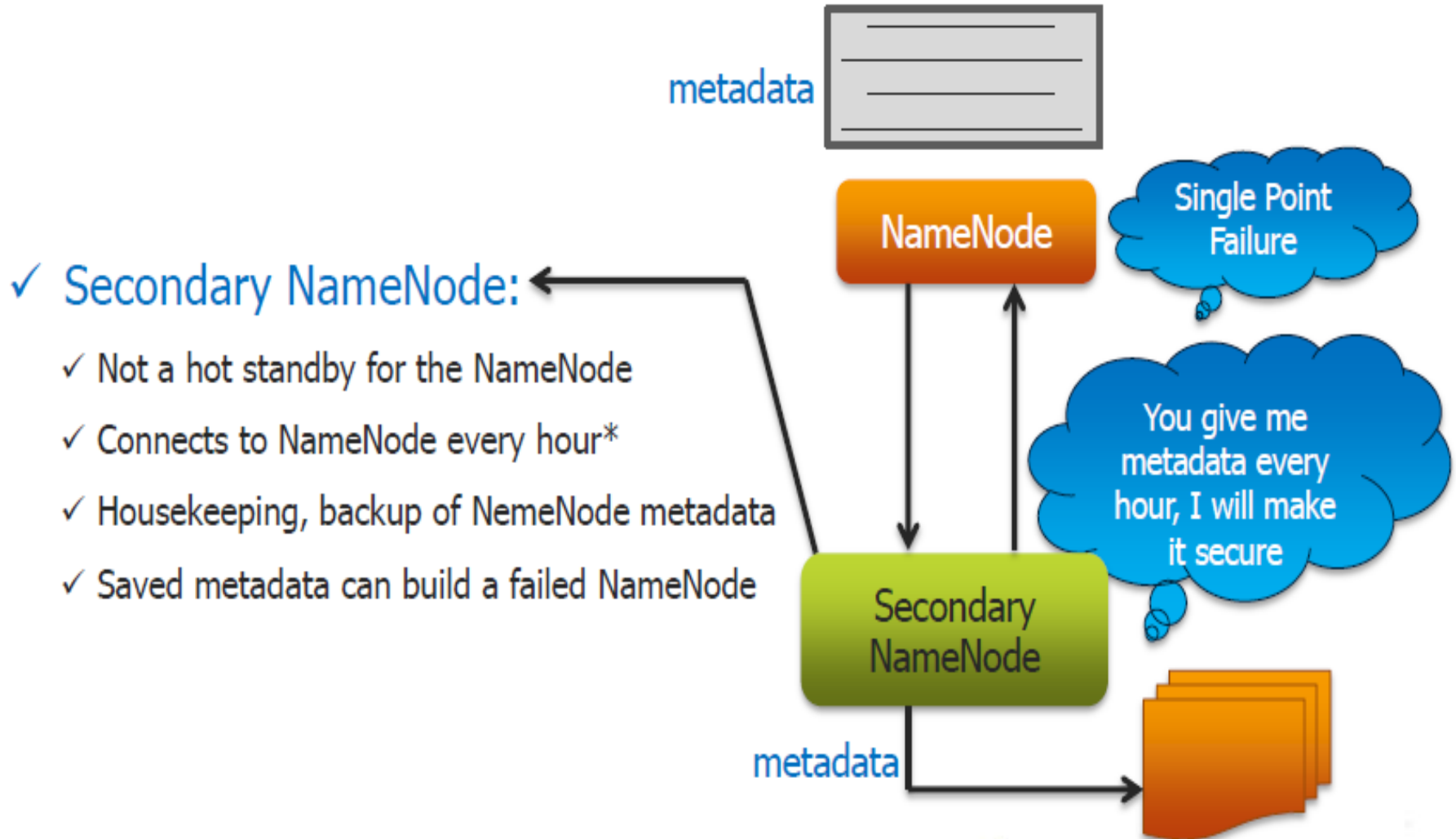


## ✓ DataNodes:

- ✓ slaves which are deployed on each machine and provide the actual storage
- ✓ responsible for serving read and write requests for the clients

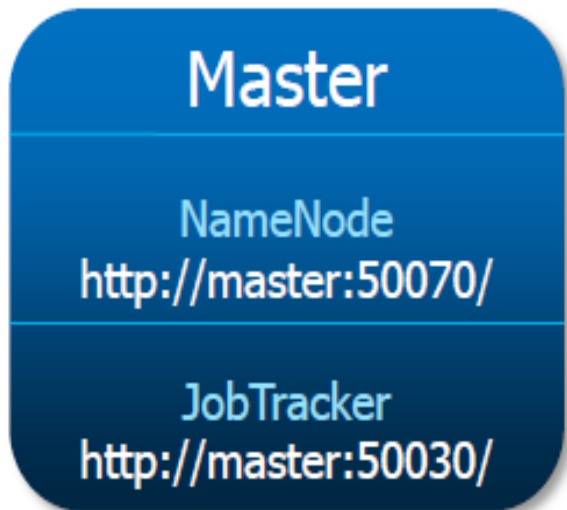


# Secondary Namenode

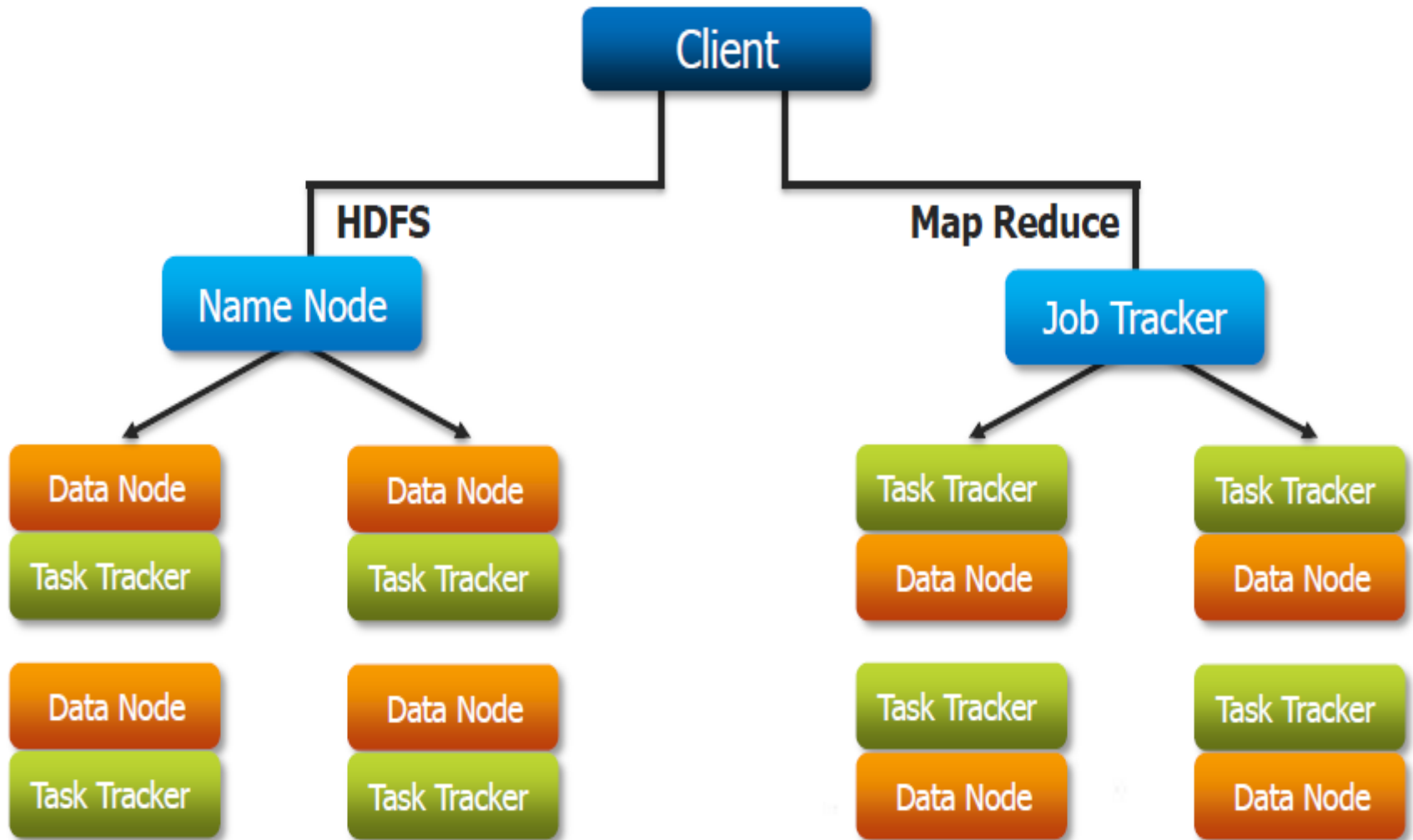




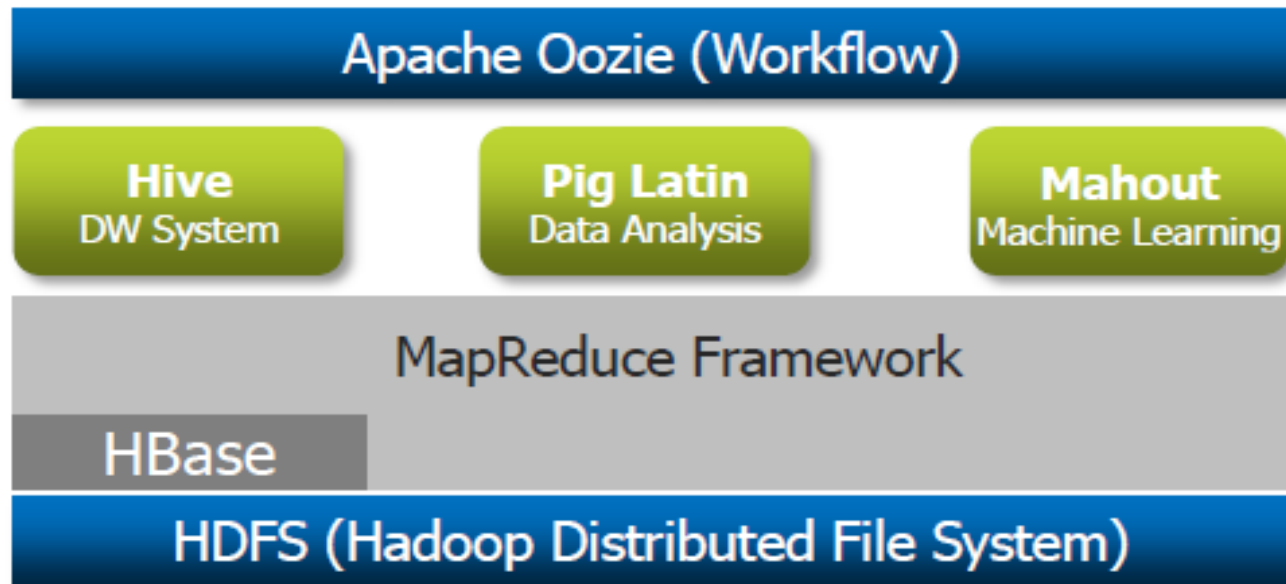
# Hadoop Cluster Architecture



# Hadoop Cluster Architecture



# Hadoop Ecosystem



Import Or Export



**Unstructured or  
Semi-Structured data**



**Structured Data**



# Hadoop Cluster Modes

Hadoop can run in any of the following three modes:

## Standalone (or Local) Mode

- ✓ No daemons, everything runs in a single JVM.
- ✓ Suitable for running MapReduce programs during development.
- ✓ Has no DFS.

## Pseudo-Distributed Mode

- ✓ Hadoop daemons run on the local machine.

## Fully-Distributed Mode

- ✓ Hadoop daemons run on a cluster of machines.

# Nowadays...

The image shows a screenshot of the Yahoo! homepage from around 2008. Several blue oval callouts are overlaid on the page, highlighting specific areas and their associated machine learning applications:

- Machine Learning (e.g. Spam filters)**: Points to the "MY FAVORITES" sidebar on the left.
- Content Optimization**: Points to the main news article titled "Miss California holds back tears".
- Search Index**: Points to the "TOP SEARCHES" list on the right.
- Ads Optimization**: Points to a Toyota advertisement for a "PERFECT TIMING EVENT".
- Content Feed Processing**: Points to the "SPOTLIGHT" section featuring "Most emailed news photos".

The Yahoo! homepage layout includes a top navigation bar with links to Web, Images, Video, Local, Shopping, and More. A search bar with a "Web Search" button is located below the navigation bar. The left sidebar contains "MY FAVORITES" and "RECOMMENDED" sections. The main content area features a large news article, a "TOP SEARCHES" list, and various other sections like "SPOTLIGHT" and "Dow" stock information.

- When you visit yahoo, you are interacting with data processed with Hadoop!

# Existing Hadoop Customers



## Retail

- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization



## Advertising & Public Relations

- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition



## Financial Services

- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis



## Media & Telecommunications

- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention



## Manufacturing

- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization



## Energy

- Smart Grid
- Exploration

## Tennessee Valley Authority Lakes



## Government

- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics



## Healthcare & Life Sciences

- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

GE Healthcare







# Hadoop vs Spark

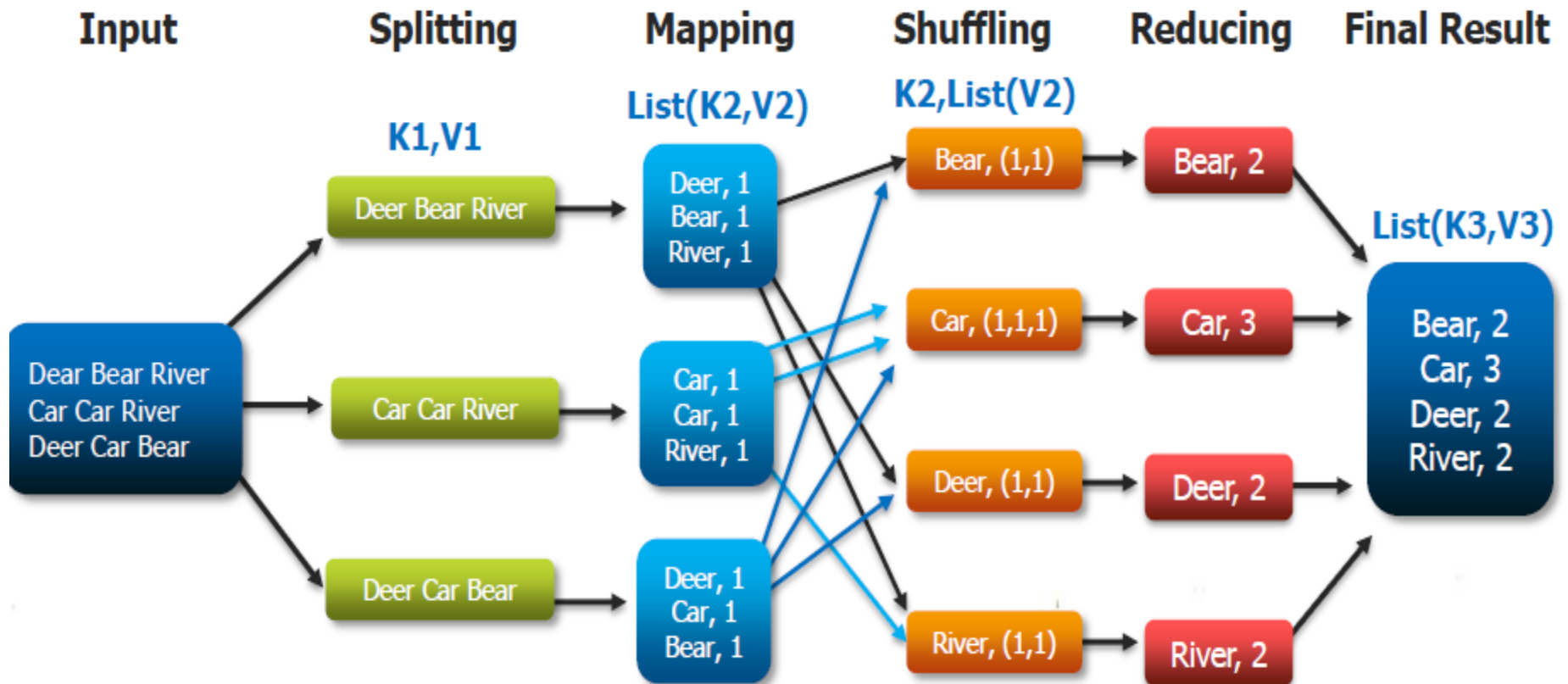
HADOOP	SPARK
Stores data on disk	Stores data in memory (RAM)
Commodity hardware can be utilized	Need high end systems with greater RAM
Uses Replication to achieve fault tolerance	Uses different data storage models to achieve fault tolerance (Eg. RDD)
Speed of processing is less due to disk read write	100x faster than Hadoop
Supports only Java & R	Supports Java, Python, R, Scala etc. Ease of programming is high.
Everything is just Map and Reduce	Supports Map, Reduce, SQL. Streaming etc
Data should be in HDFS	Data can be in HDFS, Cassandra, Hbase or S3. Runs on Hadoop, Cloud, Mesos or standalone

- <https://www.mentimeter.com/>

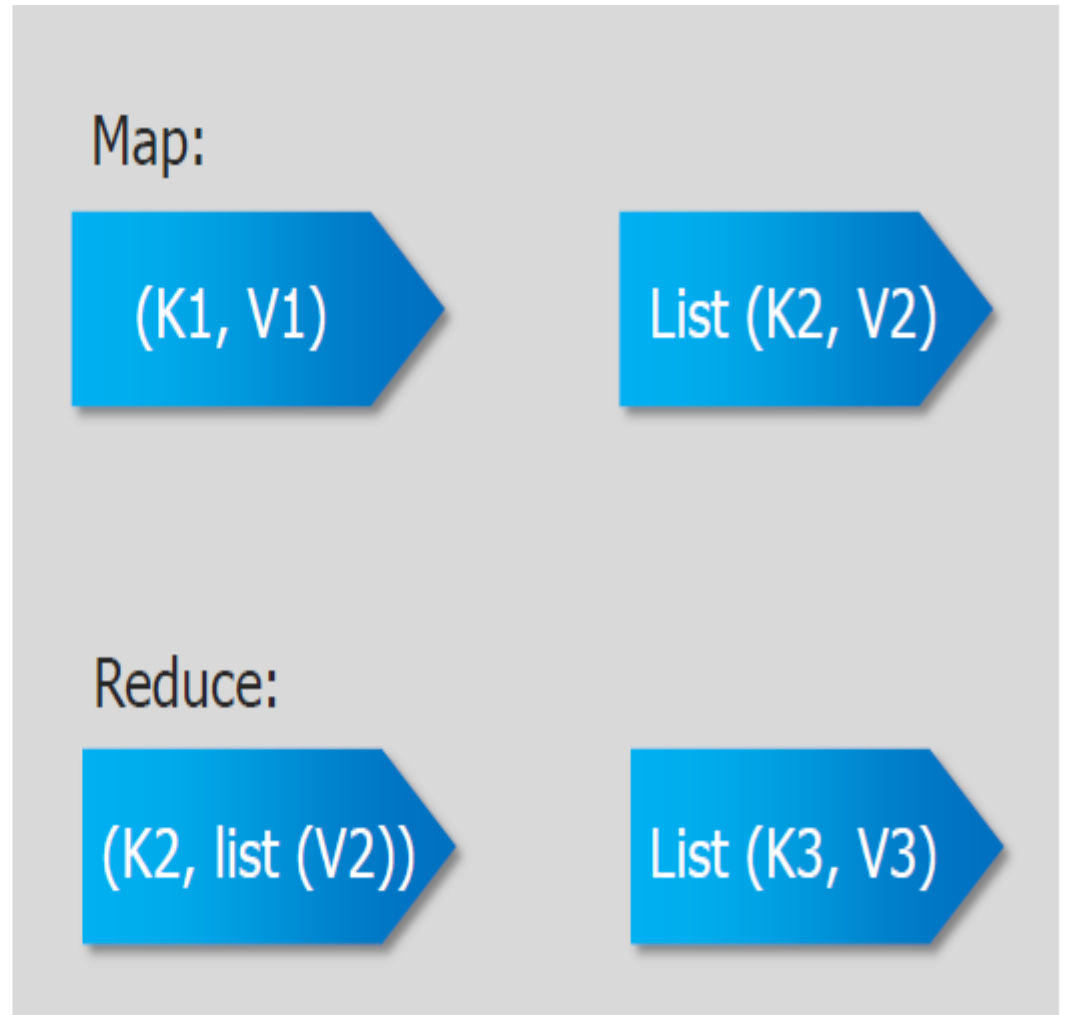
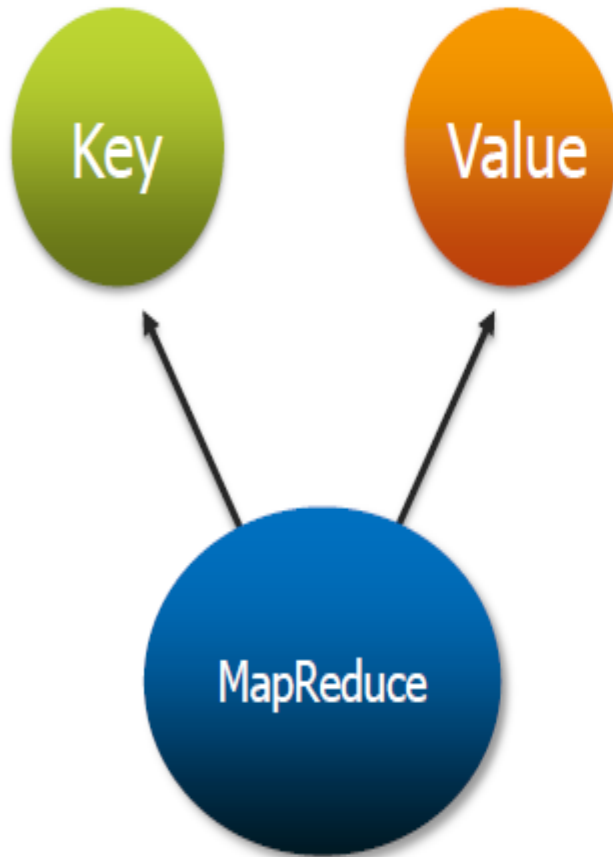


# Map Reduce Programming

## The Overall MapReduce Word Count Process



# Anatomy of a MR Program



# Map Reduce Flow

Input data is distributed to nodes

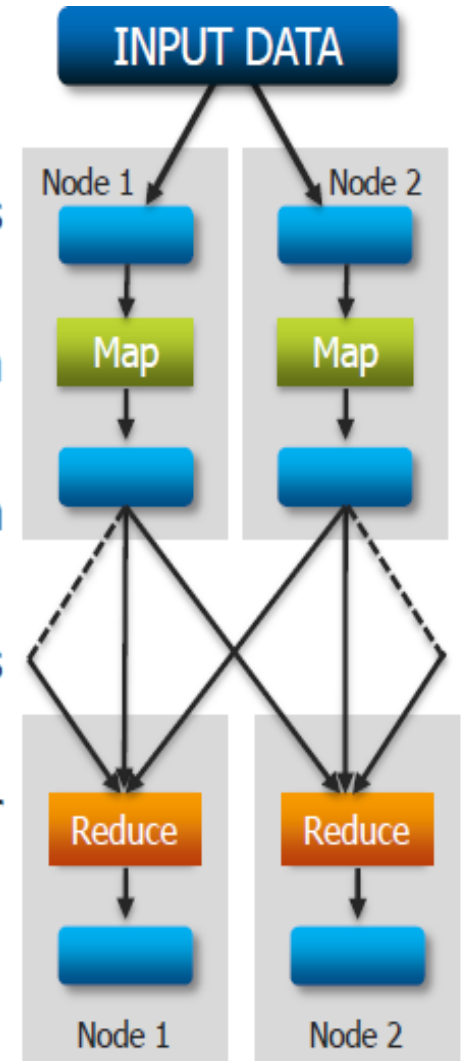
Each map task works on a “split” of data

Mapper outputs intermediate data

Data exchange between nodes in a “shuffle” process

Intermediate data of the same key goes to the same reducer

Reducer output is stored



# Annie's Question

The output of a MR job will be stored on HDFS:

- TRUE
- FALSE



# Annie's Answer

True. It is stored in different part files for eg – part-m-00000, part-m-00001 and so on. The part files are created on the basis of the block size.



# Annie's Question

To run MR job data should be present on HDFS:

- TRUE
- FALSE



# Annie's Answer

True. In order to process data in parallel it is necessary that it is present on HDFS so that MR can work on chunks of data in parallel.



# Slides Credit

- Lija Mohan, CUSAT